S1.2

A REAL-TIME IMPLEMENTATION OF THE IMPROVED MBE SPEECH CODER *

Michael S. Brandstein, Peter A. Monta, John C. Hardwick, and Jae S. Lim

Research Laboratory of Electronics
Massachussetts Institute of Technology
Cambridge, MA 02139

Abstract

This paper presents a real-time, single DSP chip implementation of a 2.4, 4.8, and 8.0 kbps improved Multi-band Excitation (IMBE) vocoder. The IMBE vocoder is based on the MBE speech model, and it has been shown to generate high quality speech under both clean and noisy conditions. In addition, the IMBE vocoder is well suited for real-time implementation since it does not require excessive computation or storage. Full-duplex operation has been demonstrated using a single AT&T WE DSP32. In this paper we will address aspects of the hardware architecture, algorithm implementation, and system performance.

1 Introduction

There has been considerable interest in the development of low bit rate, high quality speech analysis/synthesis systems. Applications for such systems include voice mail, low bit rate digital communications, and high security telephony. One class of speech analysis/synthesis systems (vocoders) which has been studied extensively and used widely in practice is based on an underlying model of speech. For this class, segments of speech are represented as the product of excitation and system spectra. The excitation parameters generally consist of a pitch period and a voiced/unvoiced (V/UV) decision. The system parameters are typically the spectral envelope or impulse response of the vocal tract. Speech is generated in the vocoder by exciting the system with a periodic impulse train in the case of voiced speech or random noise in the case of unvoiced speech. While vocoders of this type are capable of producing intelligible speech, they have not been successful in synthesizing high quality speech. In addition, the performance of these vocoders is known to degrade rapidly in the presence of background noise. Considerable attention has been devoted to improving these systems. These improvements have focused primarily on the specification and quantization of the excitation signal after removal of the pitch structure. While these techniques have improved the quality, they have significantly increased algorithm complexity, which has precluded the real-time implementation of these systems on low cost architectures.

In the Multi-Band Excitation (MBE) Speech Model a different approach is taken toward representing the excitation signal [2]. The MBE speech model replaces the binary voiced/unvoiced classification with a series of such decisions over harmonic intervals. This added degree of freedom allows each speech segment to be partially voiced and partially unvoiced. The result is a speech analysis/synthesis system which is capable of generating high quality speech in a wide range of environments without a marked increase in computational complexity.

Previous work has shown that the MBE speech model produces high quality speech at 4.8 kbps [4,11]. An improved version of the MBE model (IMBE) has recently been developed [5] which includes higher speech quality along with reduced computational requirements. An additional effort has produced a 2.4 kbps speech coding system based on the IMBE speech model. Informal listening tests have shown the new 2.4 kbps system to be substantially better than the government standard LPC-10e speech coder. The computational simplicity of the IMBE algorithm makes it particularly well suited to real-time implementation at a significantly lower cost than LPC based systems producing similar speech quality. The current system performs the full-duplex IMBE algorithm on a single WE DSP32 processor.

In the next section, we review the MBE speech model. In section 3, we briefly discuss the algorithm features. In section 4, the system hardware and software are described. In section 5, we present our implementation results.

2 Multi-Band Excitation Speech Model

Over a short-time interval, the Fourier transform $S_w(\omega)$ of a windowed speech segment $s_w(n)$ is modeled as the product of a spectral envelope $H_w(\omega)$ and an excitation spectrum $E_w(\omega)$. As in many simple speech models, the spectral envelope is a smoothed version of the original speech spectrum. The excitation spectrum in this new speech model differs from previous models in one major respect. In previous models, the excitation spectrum is totally specified by the fundamental frequency and a voiced/unvoiced decision for the entire spectrum. In this new model, the excitation spectrum is specified by the fundamental frequency and a

*This work has been supported in part by the Rome Air Development Center under Contract No. F19628-89-K-0041, and in part by a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsors.

CH2847-2/90/0000-0005 $1.00 © 1990 IEEE

voiced/unvoiced decision for each group of harmonics of the fundamental.

The excitation spectrum $E_w(\omega)$ is obtained from the fundamental frequency and the voiced/unvoiced information by combining segments of a periodic spectrum $P_w(\omega)$ in the frequency regions declared voiced with segments of a random noise spectrum in the frequency regions declared unvoiced. The periodic spectrum $P_w(\omega)$ is completely determined by the fundamental frequency. The V/UV information allows us to mix the harmonic spectrum with a random noise spectrum in a frequency-dependent manner. This model is motivated by the observation that spectra in mixed voicing segments of clean speech or in voiced segments of noisy speech tend to have regions of the spectrum dominated by harmonics of the fundamental and other regions dominated by noise-like energy. We hypothesize that humans can discriminate between frequency regions dominated by harmonics of the fundamental and those dominated by noise-like energy and employ this information in the process of separating voiced speech from random noise. Elimination of this acoustic cue in vocoders based on simple excitation models may help to explain the significant intelligibility decrease observed with these systems in noise [6].

As previously stated the new speech model assigns each group of harmonics to be either voiced or unvoiced. In [1,3] a voiced/unvoiced decision was made for each individual harmonic. However, in [2] the voiced/unvoiced information was reduced to a single decision for each group of three harmonics. This change was found to preserve the high quality capability of the MBE speech model, while substantially reducing the number of bits required to represent the voiced/unvoiced information. Using this approach noisy regions of the excitation spectrum are represented using one bit for each group of three harmonics. This is a distinct advantage over simple harmonic models [7], where noisy regions are synthesized from the coded phase requiring several bits per harmonic.

3 IMBE Algorithm

Figure 1 is an outline of the IMBE algorithm. The parameters of the MBE speech model consist of the fundamental frequency, voiced/unvoiced information, and the spectral envelope. Our approach to estimating these parameters is similar to the one presented by Griffin [2]. This approach attempts to estimate the excitation and system parameters which minimize the distance between the original and synthetic speech spectra. Rather than attempting to optimize simultaneously over all the parameters, we instead minimize the error distance over the fundamental frequency and spectral envelope assuming all voiced speech. Once these parameters are estimated, voiced/unvoiced decisions are made by comparing the spectral error over a series of harmonics to a prescribed threshold. This can be viewed as an analysis-by-synthesis system. For a given fundamental frequency the spectral envelope can be represented by a set of harmonic coefficients which correspond to the value of the envelope at the harmonics of the fundamental frequency.

Separate techniques are used to synthesize the voiced and unvoiced speech from the estimated model parameters. Voiced speech is generated as the sum of a series of sine waves. For a frame of speech, a distinct "oscillator" is assigned to each voiced harmonic. To preserve interframe continuity, the amplitude and frequency of the oscillator are interpolated between frames.

Unvoiced speech is synthesized in the spectral domain. The spectrum of a windowed noise sequence is generated and weighted by the unvoiced harmonic magnitudes. Regions corresponding to voiced harmonics are zeroed out. The inverse transform is then calculated and used with the overlap-add procedure [8] to generate the unvoiced segment. The voiced and unvoiced contributions are then added to produce the final synthesized speech.

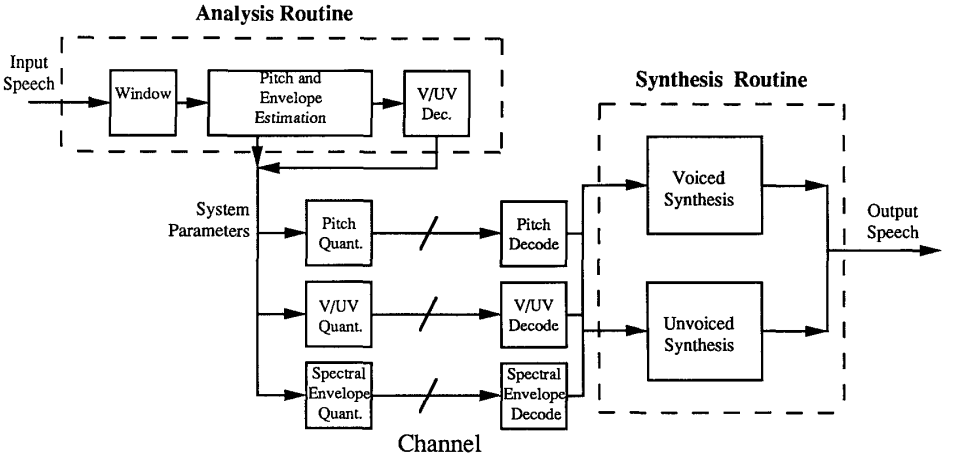

Figure 1: IMBE Algorithm Block Diagram

The model parameters which are estimated for each frame include the fundamental frequency, the voiced/unvoiced decisions, and the spectral envelope information. The number of bits available for encoding these values is a function of the bit rate and analysis frame rate. For each vocoder system implemented, 2.4, 4.8, and 8.0 kbps, the model analysis and synthesis routines are identical. With the exception of the bit allocation, the coding schemes are also similar. The fundamental frequency needs accuracy of about 1 Hz. The present encoding scheme is straightforward and requires about 9 bits per frame. The V/UV decisions are encoded with one bit per decision. The remaining bits are allocated to error control and the spectral envelope information. Experimental results demonstrate that substantial interdependencies of these magnitudes exist in both time and frequency. We adopted a transform coding approach to exploit this [4]. Through adaptive bit allocation and uniform quantization of the Discrete Cosine Transform (DCT) coefficients, a high degree of coding efficiency is obtained.

4 System Overview

Figure 2 is a block diagram of the hardware. A single WE DSP32 processor [9] is responsible for all the signal processing. The Mac II provides the user interface, a means of downloading software, and data transfer.

The DSP32 is a 32-bit floating point processor with 160 ns instruction cycle time. Its architecture and I/O capabilities make it particularly attractive for this application. A floating point processor eliminates the need to convert the IMBE algorithm to fixed point arithmetic and prevents speech quality degradations due to limited dynamic range. On-chip memory includes 2 kbytes of ROM and 4 kbytes of RAM with an additional 56 kbytes available through off-chip expansion. Distinct control and data arithmetic units provide reduced program overhead and greater computational throughput. Serial and parallel I/O ports with DMA simplify external interfacing.

In an effort to reduce design time and production cost, a commercially available DSP32 development board was purchased. The MacDSP, available from Spectral Innovations[10], sits on the Mac II Nubus and offers a DSP32 processor with 64 kbytes of external RAM, I/O control, A/D and D/A converters, and programmable interrupt capability. The board's built-in switched-capacitor filters were noisy, so we replaced them with 8th order Chebyshev filters. A number of high level Macintosh driver calls are available for effecting DMA operation and accessing on-board control registers and memory. These provide a simple means for downloading code, controlling the board, and interactively manipulating algorithm parameters.

Conversion of the IMBE algorithm to the DSP processor was accomplished with the aid of AT&T's DSP32 C compiler and simulator. The C compiler includes an assembly-level code optimizer and a number of generic and DSP libraries. The simulator is a flexible way to interactively debug and evaluate assembly level code. For efficiency, it was necessary to hand compile various sections of code. This produced code efficiency improvements by a factor of 2 to 10 for the regions involved.

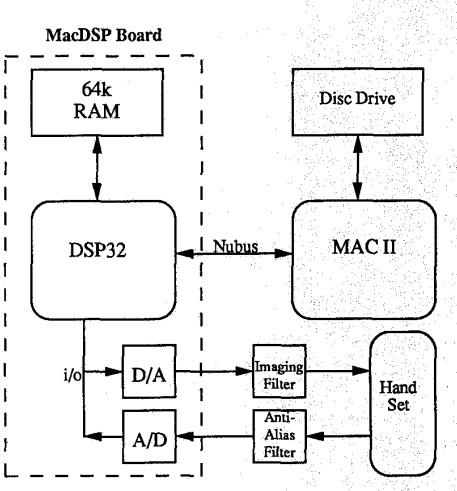

Figure 2: Diagram of System Hardware

Routine	Average Time	Maximum Time
Analysis	.42	.48
Quantization	.12	.16
Synthesis	.30	.34
Total	.84	.98

Table 1: 4.8 kbps IMBE Coder Timing as a Fraction of Real-Time on DSP32

We currently have two versions of the system available, the DSP32 based design just described and a similar DSP32C configuration. The DSP32C offers a number of enhancements over its predecessor: 80 ns instruction cycle time, a 24-bit address space, and an extended instruction set. Spectral Innovations also offers a DSP32C development board with 128k RAM. We have been using this second system for algorithm development and improvement. With twice the memory and clock speed, manual code optimization is minimized for achieving real-time operation. The turnaround time from C code software to a working vocoder is reduced to just a few minutes. Algorithm improvements may then be ported to the slower device.

5 Implementation Results

Table 1 shows a breakdown of the analysis, quantization, and synthesis routines of the 4.8 kbps coder as a function of real-time operation on a DSP32 processor. Similar figures exist for the 2.4 and 8.0 kbps implementations. In the worst case, the algorithm is running at 98 percent of real time. These results were achieved with no compromises in algorithm complexity, but did require extensive hand coding. We estimate that only 2.5 MIPS are necessary for algorithm implementation while the remainder of the 6.25 MIPS available are being expended on overhead and compiler inefficiency.

Algorithm	MIPS
IMBE	2.5
Motorola VSELP	6.4
DoD CELP	9.3
AT&T SELP	9.4
UCSB VAPC	3.9
Motorola RELP	6.2
Govt. Standard	5.4-12.5

Table 2: Computational Estimates of Some 4.8 kbps Coders

Table 2 is offered as a means for comparison of the instruction counts of various 4.8 kbps speech coders. These figures are all self reported computational estimates [11] and should not be confused with DSP chip MIPS ratings. It has been found that DSP chips require two to three times these estimated MIPS values due to program overhead. This figure agrees with our experiments. Note that none of these other systems have the potential to be implemented on the 6.25 MIPS DSP32 or an equivalent processor without some modification. Of particular interest is DoD's proposed federal standard 4.8 kbps voice coding system. This system has an upper bound of 12.5 MIPS when limited to integer pitch delays and a lower bound of 5.4 MIPS when the stochastic code book is reduced by a factor of eight [12].

In a recent government study of 4.8 kbps speech coders [11], the DAM and DRT scores for the MBE algorithm were shown to be comparable to a number of other systems, including the proposed government standard CELP system. The IMBE coder has been developed subsequent to this study and possesses a noticeable increase in speech quality over its predecessor. Although we have not performed formal intelligibility and quality tests at this point, informal listening tests by a number of subjects have placed the IMBE algorithm on par, if not superior, to other 4.8 kbps systems. We have also found that the algorithm output quality degrades gracefully with bit rate. The 2.4 kbps coder currently available sounds quite similar to its 4.8 kbps counterpart.

6 Conclusions

We have presented an implementation of the MBE speech coder in real time, on commercially available hardware. The computational simplicity and high quality speech production of the IMBE algorithm make it a practical candidate for low bit rate speech coding applications.

References

[1] Daniel W. Griffin and Jae S. Lim, "A New Model-Based Speech Analysis/Synthesis System," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 513-516, Tampa, Florida, March 26-29, 1985.

[2] Daniel W. Griffin, "Multi-Band Excitation Vocoder," *Ph.D. Thesis*, E.E.C.S. Department, M.I.T., 1987.

[3] Daniel W. Griffin and Jae S. Lim, "A High Quality 9.6 kbps Speech Coding System," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 125-128, Tokyo, Japan, April 13-20, 1986.

[4] John C. Hardwick and Jae S. Lim, "A 4.8 KBPS Multi-Band Excitation Speech Coder," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 374-377, NY, NY, April 11-14, 1988.

[5] John C. Hardwick and Jae S. Lim, "A 4800 bps Improved Multi-Band Excitation Speech Coder," *IEEE Speech Coding Workshop*, Vancouver, B.C., Canada, Sept. 5-8, 1989.

[6] B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," M.I.T. Lincoln Laboratory Technical Report, TR-670, December 1983.

[7] R. J. McAulay and T. F. Quatieri, "Mid-Rate Coding Based on a Sinusoidal Representation of Speech," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 945-948, Tampa, Florida, March 26-29, 1985.

[8] Daniel W. Griffin and Jae S. Lim, "Signal Estimation From Modified Short-Time Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236-243, April 1984.

[9] *WE DSP32 Digital Signal Processor Information Manual*, AT&T, 1988.

[10] *MacDSP User's Manual*, Spectral Innovations,Inc., 1989.

[11] D.P. Kemp, R.A. Sueda, T.E. Tremain, "An Evaluation of 4800 BPS Coders," *Proc. of the Military and Government Speech Tech '89*, pp. 86-90, Arlington, VA, Nov. 13-15, 1989.

[12] Joseph P. Campbell,Jr.,Vancy C. Welch, and Thomas E. Tremain, "The New 4800 bps Voice Coding Standard," *Proc. of the Military and Government Speech Tech '89*, pp. 64-70, Arlington, VA, Nov. 13-15, 1989.